2nd Project portfolio

Introduction

Available on Kaggle is a “Sample Sales Data” set which is quite rich and includes information on orders and order details, sales and customer data as well as shipping data. This makes it a very useful tool to use retail analytics tasks for instance segmentation, customer analysis and clustering. This dataset is especially useful when making analysis with the aim to understand customer’s actions and improve the performance of the enterprise. The information it provides is on almost every country in the world, giving it a strong position of being used in the formulation of viable strategies to grow the sales and customer satisfaction sectors of e-commerce.

Data Cleaning and uses of Software:

At a first step of the analysis, I load the data into the Python environment and perform initial data preparation. This includes features such as deletion of redundant features, dealing with missing values,

normalization of data types and verification of data consistency using the Pandas and NumPy of Python. In data analysis, I use Python queries to analyze the cleaned data and find out insights and to solve problems as well. For visualizations, I make use of the Python visualization libraries such as Matplotlib and Seaborn; also, I create a detailed Power BI report containing visualizations to improve the investigation of the results.

Data Analysis and EDA:

I began with Exploratory Data Analysis (EDA), a crucial step in the data analysis process. EDA involves systematically examining the dataset to uncover its key characteristics, identify patterns.

- Importing libraries: importing python libraries like Numpy, pandas, Matplotlib and seaborn and sns.set function
- Loading Data: the data is loaded Using pandas libraries

Data cleaning:

- The data is check by null function to fi d the null values.
- Checking Data frame shape and column 2823 rows and 25 column
- Checking Data type and correcting by code converting Data into datatype (data column into datatype)
- Remove duplicate and irrelevant Data
- Merged names column into one column contact names
- Checking the data describe mean, mode and standard deviation of the data
- Filling null values and rename the columns

Correlation Analysis: correlation analysis between the variable which identify the relation between the variables by Range 1 -1 and 0. Which interpret the positive, Negative and no correlation.

Data Dictionary:

Variable	Defination
1. ORDERNUMBER	Unique identifier for each order.
2. QUANTITYORDERED	Number of items ordered.
3. PRICEEACH	Price of each item.
4. ORDERLINENUMBER	Sequence of the order line.
5. SALES	Total sales amount.
6. ORDERDATE	Date of the order.
7. STATUS	Current status of the order.
8. QTR_ID	Quarter identifier.
9. MONTH_ID	Month identifier.
10. YEAR_ID	Year identifier.
11. PRODUCTLINE	Category of the product.
12. MSRP	Manufacturer's suggested retail price.
13. PRODUCTCODE	Code of the product.
14. CUSTOMERNAME	Name of the customer.
15. PHONE	Customer’s contact number.
16. ADDRESSLINE1	Primary address line.
17. ADDRESSLINE2	Secondary address line.
18. CITY	Customer’s city.
19. STATE	Customer’s state.
20. POSTALCODE	Customer’s postal code.
21. COUNTRY	Customer’s country.
22. TERRITORY	Sales territory.
23. CONTACTLASTNAME	Last name of the contact person.
24. CONTACTFIRSTNAME	First name of the contact person.
25. DEALSIZE	Size of the deal.

Question and Insights

Question No 1identifity the number of sale over past serval years

Insight:

The change in sales was very large between 2003 and 2004, While the record of 4 was set in 2004. 4.7M in sales implying an uptrend year on year sales growth. The maximum sales recorded for \2004 relatively to 2003 and 2015 can be referred to as the sales peak which may observed in association with successful product launches, market expansion or due to sales peaks in certain seasons. Wherein EMEA region again had the highest sales throughout the year due to good market sales and NA and APAC were trailing behind.

Question No 2 Which Month Has the Most Sales?

Insights

A marked rise in sales in the 11th month can be attributed to the consumption patterns related to the fascination related to year-end sales while the other months record low sales, it provides a clear indication of the fact that November can be critical period to target promotions and sales. Other months show smaller sales numbers, which means that November is an important time for business to achieve maximum sales and concentrate on advertising.

question:3 Which country have most sale

Insights

The United States dominates sales, contributing over 52% of total global sales, indicating it is the primary market driver while the France and Spain is following by USA.

Question 4: How does the total sales distribution vary across different territories on a month-by-month basis?

Insights:

Seasonal sales are at their highest in November in all regions and are considerably higher in EMEA and North America. EMEA region remains almost constant with respect to its sales throughout the year where as in North America it goes up in May and November. APAC and Japan have more fluctuating sales changes.

The analysis of the sales distribution throughout the months shows that November records the highest sales all through the year in all the territories. North American region has higher sales rate because it has a large consumer market while the Japan has the lowest sales rate maybe because of different purchasing behavior or market congestion.

Question:5 What are the total sales by quarter?

Insights: The distribution by quarters shows that Q4 is by far the most productive, this is mainly because there are many large orders before the end of the year, many businesses require many cars, every company plans its expenses for the next year and aims to spend most of it before the end of the year. This is near to the truth that are shown by sales trends: super high in Q1 because it marks beginning of the year, sales are low in Q2 and fairly acceptable in Q3.

The histogram below shows the quarterly sales in 4 different shades of blue that represents each quarter of the year. The height of each bar represents how often particular sales amounts fall into specific categories, and the quarters are coloured differently.

The vertical red dashed line identifies the peak sales point more generally in the last quarter (shaded darkest blue). This line gives the median of the frequencies, and is used to show the highest point of sales density during the period of Quarter 4.

***Question:6 define the most order product type ***

Insights:

The above graph shows that the products with the lowest price were sold the most, whereas the products with a high price were the least popular.

From the plot, 'Classic Cars" & "Vintage Cars" are the most demanded products, 'Trains" was the least demanded product.

***Question 7 :Calculate the average sale per deal size ***

Insights: The average sale per deal size shows a clear trend: Thus, it suggests that buying through the identified channels result in significantly higher revenues given Large deals yields significantly more revenues than Medium and Small deals. This means that increasing emphasis on high value business has a potential to significantly increase the total turnover. The trend showed that higher numbers of large deal sizes, respectively EMEA and USA regions are more lucrative as medium and small deal sizes constitutes lesser incremental sales.

Question **8: find the Distribution of Deal Sizes Across Different Territories?

Insights: EMEA has the largest total number of large deals suggesting a robust market for large transactions and also displays reasonable counts of medium and small sized transactions. North America has a strategic and well-developed market containing many middle and small transactions, but less large one than EMEA. Out of all the regions and countries, both APAC and Japan demonstrate smaller numbers of large deals, with Japan having the smallest average deal amounts by a wide margin, so there may be opportunities for market expansion.

Q**uestion 9: What are the counts of deals for each product line and deal size, and how do these counts vary across different product lines? bold text**

Insights: Further analysis shows that Medium size Deals are most popular with Classic Cars and Vintage Cars while in the small deals, Classic Cars stand high. Classic car offers a high rate in USA a market. Motorcycles and Planes have less number of product categories in Large Deal, this

imply that these products are used in few large value selling transactions. Trains and Trucks and Buses are categorized as less in deal totality, especially trains are involved less in high valued deals, and therefore it can be a sign of new opportunities for development or reassessment on such sales approach.

Question 10: How do total sales differ across various order statuses

Insights: The total sales differ across order status where Shipped order produced the most sales perhaps as a pointer that the order has been completed or fulfilled. It turns out that Cancelled and Disputed statuses indicate significantly lesser sales, which are evidence of lost profits or the presence of problems. The rest statuses, including Process, On Hold, and Resolved, are in between, which means that they help to make sales, but they do not have the same influence as completed shipments. This strikes the need to ensure that there is a strong focus in order fulfillment as it augments the sales revenue.

Question 11: How are order statuses distributed across different territories, and what does the distribution reveal about the status of orders in each territory?

Insights: In the geographical context, two regions of EMEA and North America are showing the highest levels in shipped orders, thus signalling good delivery performance. EMEA also has more numbers in the category of cancelled and disputed orders indicating problems with order order processing. North America has rather a large number of orders on ice, which may hint at delivery issues. The number of orders shipped from Japan is above average and a small percentage of them experience some problems while most of the orders are processed efficiently.

Finding the correlation Analysis between variables

Insights:

- **Correlation Coefficient**: Each cell in the matrix contains a number between -1 and 1, representing the correlation coefficient between two variables.
- ORDERNUMBER and ORDERDATE have strong influence and they are positively related up to a high degree, 0. 98 to be precise, this suggest that as the order number increases, it is also

possible to order later in date. This makes sense as order numbers more often than not is an incrementing number to reflect the time series of the orders.

- QUANTITYORDERED and SALES: These variables also have direct relationship of 0. 55, which show that, the more the quantity ordered the more the total sales amount will also be.
- PRICEEACH and SALES, PRICEEACH and MSRP, YEAR_ID and ORDERNUMBER, QTR_ID and MONTH_ID. They have a positive correlation.
- This matrix is quickly identify the relationship between variables which can help in modeling.

Recommendations for stakeholders

1. Sales trend analysis: Inspect several years' sales data to evaluate how sales have performed over a specific period. Compare trends from different years to identify long-term sales changes caused by economic conditions, the market environment, and sales promotion strategies.
2. Monthly Sales Analysis: Monthly sales report is meant to assist the business to know the monthly sales of the business. And will assist in analyzing more and to determine more the factors which contributed to the increase or decrease of sales in some months. They can also be used to organize future activities like having promotion or marketing activities that are relevant to the month.
3. Focusing on customers in high-sales locations: Concentrating on the customers in the high-sale areas is also likely to boost sales in a better way. Each of such market area will require a set of marketing strategies and promotional activities.
4. The best-selling products are the cars. The EMEA Territory has the biggest sales, with the USA contributing the highest sales per country.
5. Improving Products or Services: Studying the products or services usually required by the clients can enable the organization identify changes in customer needs. For enhancing or creating the products or services more effectively to satisfy the needs of the customers.
6. Quarter 4, especially November and October, showed higher sales, likely due to anticipation of the December festive period. Increasing my marketing efforts in quarter 4 will likely result in higher sales.

7. Best-selling product rankings: Rankings of the Best-Selling product may be used to identify and allocate resources to those products with the greatest promise on the market. It also aids in enhancing the visibility of such goods and to get those goods to the market faster.
8. Implement improvements in order processing to reduce high cancellation and dispute rates in EMEA. Streamline procedures and resolve logistical challenges in North America to enhance overall fulfillment efficiency and customer satisfaction.
9. Well the bestselling products are the cars and the most lucrative territory is the EMEA territory considering the number of countries to target, it will be best marketing strength is targeted towards the US which us far more lucrative. So targeting The USA market with our cars will be a good marketing strategy.